Consistency of a Trained Taste Panel

NAOMI E. LIMING
Eastern Utilization Research and Development Division, USDA, Washington, D. C.

Abstract

A means of determining the reliability of a trained taste panel is to test their consistency of judgments on duplicate samples. Fourteen hundred and twenty-three duplicate sets of research milk samples were tested by 19 judges, for identification and measurement of intensity of off-flavor. The samples were scored on a 10-point scale ranging from 31 to 40. Each of the 17 flavor criticisms is based on a 6-point range. Analysis of data indicates an acceptable degree of consistency in panel performance, even though there is considerable variation among judges. The average absolute differences determined on a monthly basis ranged from 0.10 to 0.39 for a mean of 0.26 for the panel as a whole, whereas, for individual judges the range was 0.41 to 0.80 for a mean of 0.65.

The research program of the Dairy Products Laboratory includes the flavor evaluation of a large number of experimental milk samples for a) identification, and b) measurement of intensity of off-flavor. Since October, 1957, a trained taste panel has served as an analytical tool to provide an unbiased sensory evaluation of experimental samples of processed milks. Because the reliability of flavor evaluation depends upon the sensitivity and consistency of the judges, it is important to maintain a constant check on their performance. The objective of this paper was to determine individuals' ability to reproduce their evaluation of duplicate samples.

Experimental Procedure

One hundred and forty-five sets of duplicate milk samples were randomly scheduled among other samples, so the judges could not anticipate when the duplicates might be included. Four types of experimental milks (butteroils, whole milk powders, ion-exchange milks, and sterilized concentrates), together with fresh market milks, were included among the duplicates, to assure representation of different off-flavors. Of these duplicate sets, 45.5% were butteroils, 27.6% whole milk powders, 20.7% fresh fluid milks, 3.4% ion-exchange milks, and 2.8% sterilized concentrates. Reconstitution of the milks and panel procedures were in accordance with methods previously reported by Tamsma et al. (2).

The scoring system used was based on a score-card developed by the A.D.S.A. Score-Card Committee (1963) and is similar in many respects to that used by collegiate dairy judging teams. The scale ranges from 31 to 40. There are 17 flavor criticisms detectable, with a 6-point range in score for each criticism. More than one flavor criticism may be assigned to a sample. In this study only the predominant criticism was considered.

Results and Discussion

Fourteen hundred and twenty-three sets of duplicates were evaluated by 19 judges within a 19-month period. Of the total sets tested, 798 (56%) scored no point difference, 445 (31%) within one point, 134 (9%) within two points, 36 (2.5%) within three points, and ten (0.7%) within four points (Table 1). Four of the judges were responsible for all sets that varied by four points, and for more than half of the 36 sets that varied by three points. Six of the judges were able to reproduce judgments within one point on 90% of the duplicates they tested. Even though there was considerable variation in individual judges, the panel as a whole was able to repeat itself within one point on 87% of all samples tested.

The average absolute differences between duplicates are plotted on a monthly basis in Figure 1. The panel as a whole ranged from 0.10 to 0.39 for a mean of 0.26, and individual judges ranged from 0.41 to 0.80 for a mean of 0.65.

The flavor criticism used most frequently by all judges was stale for 1,051 (36.9%) of the samples (Table 2). This could be expected, as more than 85% of the reconstituted butteroils and whole milk powders were tested after storage. The criticisms next in order were unclean and feed. Chalky was the only flavor criticism not used.

Fourteen judges used stale most frequently, two used unclean, and one each used cooked, feed, and flat. Eight of the above 14 used feed as second in frequency, four used unclean, and

Received for publication January 3, 1966.

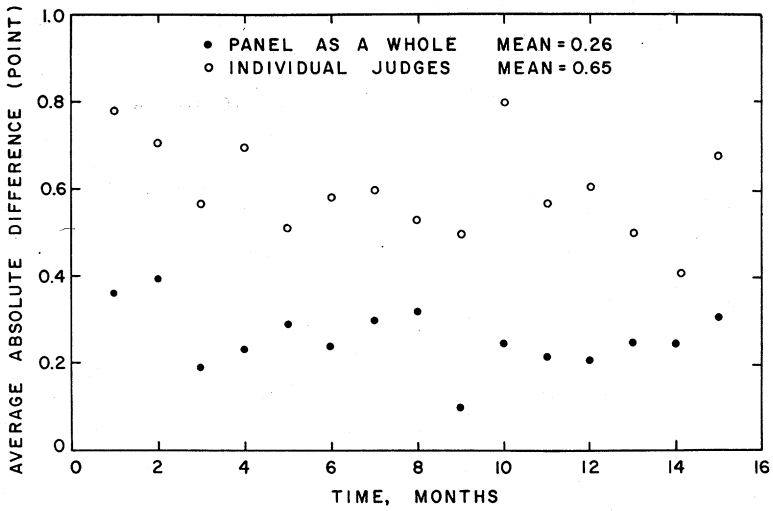

FIG. 1. Average absolute difference between duplicates.

the two others used cooked. Of the two who used unclean most frequently, stale and foreign were their second most used. The remaining three paired off with cooked first and flat second, feed first and unclean second, and flat first and stale second.

It is evident that some judges do not use certain flavor criticisms. This could indicate a low sensitivity to that particular flavor or an inability to identify it. It would appear advantageous to determine individual thresholds to the off-flavors, in order to assign judges to those panels testing the off-flavor milks to which they are most sensitive. Such a procedure is generally followed, but many times it is impossible to predict which off-flavors are present. A panel of sufficient number is necessary to assure a representative sensitivity to off-flavors of varying intensities. Kirkpatrick et al. (1) reported that a person acutely sensitive to one off-flavor in reconstituted dry milk may not be sensitive to all off-flavors. We have observed that with increased intensity of off-flavor the judges were better able to reproduce judgments and were more homogeneous in flavor identification.

Taking the panel as a whole, the whole milk powders had the least point difference for a

TABLE 1

Extent of ability of 19 judges to reproduce judgments in evaluation of duplicate samples

Judge	Sets tested No.	%	No-Point Difference No.	%	1-Point Difference No.	%	2-Point Difference No.	%	3-Point Difference No.	%	4-Point Difference No.	%	% Consistency within 1 point
I	36	24.8	28	77.8	8	22.2	.....	.....	....	....	....	....	100.0
S	70	48.3	56	80.0	13	18.6	1	1.4	....	....	....	....	98.6
G	137	94.5	84	61.3	50	36.5	3	2.2	....	....	....	....	97.8
H	115	79.3	77	67.0	32	27.8	5	4.3	1	0.9	....	....	94.8
F	114	78.6	71	62.3	33	28.9	8	7.0	2	1.8	....	....	91.2
K	75	51.7	41	54.7	27	36.0	7	9.3	....	....	....	....	90.7
D	81	55.9	34	42.0	38	46.9	7	8.6	2	2.5	....	....	88.9
P	18	12.4	11	61.1	5	27.8	2	11.1	....	....	....	....	88.9
N	118	81.4	63	53.4	40	33.9	14	11.9	1	0.8	....	....	87.3
M	15	10.3	5	33.3	8	53.3	1	6.7	1	6.7	....	....	86.6
T	107	73.8	61	57.0	31	29.0	12	11.2	3	2.8	....	....	86.0
A	121	83.4	72	59.5	31	25.6	12	9.9	4	3.3	2	1.6	85.1
C	33	22.7	17	51.5	11	33.3	4	12.1	1	3.0	....	....	84.8
J	30	20.7	16	53.3	9	30.0	4	13.3	1	3.3	....	....	83.3
B	105	72.4	59	56.2	28	26.7	14	13.3	4	3.8	....	....	82.9
O	89	61.4	33	37.1	36	40.4	12	13.5	6	6.7	2	2.2	77.5
E	103	71.0	48	46.6	30	29.1	14	13.6	6	5.8	5	4.8	75.7
R	33	22.8	14	42.4	9	27.3	6	18.2	3	9.1	1	3.0	69.7
L	23	15.9	8	34.8	6	26.1	8	34.8	1	4.3	....	....	60.9
	1,423		798	56.1	445	31.3	134	9.4	36	2.5	10	0.7	87.4

TABLE 2

Extent of use of milk flavor criticisms by 19 judges in evaluation of duplicate samples

Description	Judge: I	S	G	H	F	K	D	P	N	M	T	A	C	J	B	O	E	R	L	No. times used	% times used
Acid					X												X			6	0.2
Astringent			X		X	X									X	X		X		38	1.3
Bitter	X																X			3	0.1
Chalky																					
Cooked		X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	121	4.2
Feed	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	453	16.0
Flat			X	X	X	X	X	X	X	X		X	X	X	X		X	X		215	7.6
Lactone		X	X	X	X	X	X		X		X	X	X		X	X	X			167	5.9
Metallic					X							X								6	0.2
Oxidized	X		X	X	X	X	X		X	X	X	X	X		X	X	X	X	X	79	2.8
Rancid	X								X								X			11	0.4
Salty											X									2	0.1
Scorched			X								X									6	0.2
Stale	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	1,051	36.9
Unclean	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	530	18.6
Weedy											X									2	0.1
Foreign	X		X	X	X	X	X	X	X			X			X	X	X		X	156	5.5

mean of 0.22, and the ion-exchange milks the greatest for a mean of 0.38. The three other milk groups had a mean point difference of 0.25 for sterilized concentrates and fresh fluid milks, and 0.29 for the butteroils.

Among judges the least point difference was with the sterilized concentrates for a mean of 0.37; the greatest was the ion-exchange milks for a mean of 0.90. The other milk groups had a mean of 0.58 for fresh fluid milks and 0.60 for the whole milk powders and butteroils.

Conclusions

A planned scheduling of duplicate samples gives a good means for determining consistency of judgments for individual judges and a trained panel as a whole, and provides a basis for deciding which judges need additional training. Interpretation of the results suggests that for a trained panel to function acceptably as an analytical tool the individual judges should be able to reproduce judgments within one point (ten-point scale) on at least 80% of the samples tested. Judges who fail to show this degree of consistency should have additional training.

Acknowledgment

The author gratefully acknowledges the technical assistance of J. N. Boyd, Biometrician.

References

(1) Kirkpatrick, M. E., Lamb, J. C., Dawson, E. H., and Eisen, J. N. 1957. Selection of a Taste Panel for Evaluating the Quality of Processed Milk. Food Technol., 11:3.

(2) Tamsma, A., Pallansch, M. J., Mucha, T. J., and Patterson, W. I. 1961. Factors Related to the Flavor Stability of Foam-Dried Whole Milk. I. Effect of Oxygen Level. J. Dairy Sci., 44:1644.